INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Convergence of MCEM and related algorithms for hidden Markov random field

Nathalie Peyrard

N° 4146

Mars 2001

THÈME 4

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--4146--FR+ENG

Convergence of MCEM and related algorithms for hidden Markov random field

Nathalie Peyrard

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet is2

Rapport de recherche n° 4146 — Mars 2001 — 14 pages

Abstract: The Monte Carlo EM (MCEM) algorithm is a stochastic version of the EM algorithm using MCMC methods to approximate the conditional distribution of the hidden data. In the context of hidden Markov field model-based image segmentation, the behavior of this algorithm has been illustrated in experimental studies but little theoretical results have been established. In this paper new results on MCEM for parameter estimation of the observed data model are presented, showing that under suitable regularity conditions the sequence of MCEM estimates converges to a maximizer of the likelihood of the model. A variant of the Monte Carlo step in the MCEM algorithm is proposed, leading to the Generalized Simulated Field (GSF) algorithm, and it is shown that the two procedures have similar properties.

Key-words: Hidden Markov random field, EM algorithm, Monte Carlo simulations, Gibbs sampler, Generalized EM, Plug-in methods, Image segmentation.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Convergence de l'algorithme MCEM et de procédures similaires pour champ de Markov caché

Résumé : L'algorithme Monte-Carlo EM (MCEM) est une version stochastique de l'algorithme EM dans laquelle la distribution conditionnelle des données cachées est approximée par une méthode de type MCMC. Dans le contexte de la segmentation markovienne d'image, le comportement de cet algorithme a été illustré de manière expérimentale, mais peu de résultats théoriques sont disponibles. Dans ce rapport, de nouveau résultats pour l'estimation des paramètres de la distribution des données observées sont présentés, montrant que sous des conditions raisonnables de régularité, la suite d'estimateurs MCEM converge vers le maximiseur de la vraisemblance du modèle. Une variante pour l'étape M de l'algorithme MCEM est proposée, qui conduit à un algorithme du *champ simulé généralisé* (GSF) et il est montré que les deux procédures ont des propriétés similaires.

Mots-clés : Champ de Markov caché, Algorithme EM, simulations de Monte-Carlo, Échantillonneur de Gibbs, Algorithme EM généralisé, Méthodes de type plug-in, Segmentation d'image.

1 Introduction

In this paper, we address the problem of parameter estimation for hidden Markov random fields (HMRFs). In image segmentation, these models are useful to describe spatial dependencies between neighboring pixels. The Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) is a powerful tool for parameter estimation on incomplete data problems. When part of the data is hidden, the direct computation of the Maximum Likelihood Estimate (MLE) is usually not feasible. Rather, the EM algorithm uses the fact that the complete data distribution can be simpler to deal with, to iteratively compute the MLE. Each iteration is divided in two steps: the Expectation (E) and the Maximization (M) steps. The E step consists of computing the expectation of the log-density of the complete data, conditionally on the observed data and a current value of the parameter. In the M step, the parameter is updated by maximizing this expectation. When the model for the hidden data is a Markov random field, EM is difficult to carry out, due to the complex structure of this model. In the E step, the computation of the conditional Markov distribution is involved, and is in general not tractable. An alternative is to use Monte Carlo simulations to generate realizations of the conditional hidden field. The expectation in the E step is then replaced by the empirical mean of the complete log-density. This is the principle of the Monte Carlo EM (MCEM) algorithm, first proposed by Wei and Tanner (1990).

Little is known about the convergence properties of MCEM for HMRFs. When in addition to the parameter θ of the observed data model, the spatial parameter of the Markov distribution has to be estimated, the problem is quite complex and there is little theoretical results. In the case where the spatial parameter is known, Comer and Delp (2000) considered a stochastic version of EM, the EM/MPM algorithm, which consists of plugging estimates of the conditional distributions, resulting from a the Gibbs sampler, in the formula of the M step. Assuming that the spatial parameter is known, they established that almost surely the EM/MPM estimate of the parameter θ can be made arbitrary close to the EM estimate, if a sufficient number of iterations of the algorithm and of the Gibbs sampler are performed. It means that if the EM estimate is close to the true value of the parameter, then the EM/MPM estimate have the same nice characteristic. It provides first theoretical results on the MCEM convergence, since it is pointed out in this article that EM/MPM is equivalent to the MCEM algorithm.

In this paper, we consider as well that the spatial parameter is known, and we adopt a different point of view to establish new and more general results on the convergence of MCEM. The behavior of this algorithm is related with that of a Generalized EM (GEM) (see Dempster, Laird, and Rubin 1977), in order to use the theoretical results known about these algorithms. More precisely, conditions are given to ensure that with probability one the log-likelihood sequence is not decreased and converges. It is shown then that under suitable regularity conditions, the MCEM estimate converges almost surely to a stationary point or a maximizer of the log-likelihood of the observations. These results can be extended to any variant of the MCEM algorithm using a MCMC method to approximate the conditional probabilities of interest, as soon as the approximations converge to the true values. In particular, we prove the convergence of the Generalized Simulated Field (GSF), derived from mean field-like approximations of a Markov field (Zhang 1992), and proposed in Celeux, Forbes, and Peyrard (2001)).

The paper is organized as follow. The MCEM procedure is described in Section 2, and its links with the EM/MPM algorithm and the GSF algorithm is shown in Section 3. Convergence properties are established in Section 4. The methods studied are illustrated in Section 5, on a synthetic example of image segmentation.

2 MCEM algorithm for Hidden Markov random field

The context of this work is image segmentation. The complete data are denoted $(\mathbf{Y}, \mathbf{Z})$, where $\mathbf{Y}$ is the observed image and $\mathbf{Z}$ represents the unobserved pixels' classification. The hidden field $\mathbf{Z}$, as well as $\mathbf{Y}$, are defined on a finite set of sites, S, corresponding to the pixels of the image. If the number of classes is K, the components $Z_i,\ i \in S$ take values in $\{1, \ldots, K\}$. In Markov model-based image segmentation, the joint probability distribution of $\mathbf{Z}$ satisfies the following properties,

$$\begin{aligned} \forall\, \mathbf{z}, \quad P_G(z_i \mid \mathbf{Z}_{S\backslash\{i\}} = \mathbf{z}_{S\backslash\{i\}}) &= P_G(z_i \mid Z_{N(i)} = z_{N(i)}), \\ \forall\, \mathbf{z}, \quad P_G(\mathbf{z}) &> 0, \end{aligned}$$

where $\mathbf{z}_{S\backslash\{i\}}$ denotes a realization of the field restricted to $S\backslash\{i\} = \{j \in S, j \neq i\}$ and $N(i)$ denotes the set of neighbors of i, according to the neighborhood system defined on S. The Hammersley-Clifford theorem states that there is an equivalence between the above definition of the joint probability distribution of a Markov random field, and a Gibbs distribution,

$$P_G(\mathbf{z}) = W^{-1} \exp(-H(\mathbf{z})),$$

where H is the so-called energy function. The observed data $\mathbf{Y} = \{Y_i,\ i \in S\}$ are supposed to be independent conditionally on $\mathbf{Z}$, and the conditional distribution is of the following form

$$g(\mathbf{y} \mid \mathbf{Z} = \mathbf{z},\ \theta) = \prod_{i \in S} g_i(y_i \mid Z_i = z_i,\ \theta), \tag{1}$$

where the conditional densities g_i's are positive and depend on a parameter θ in $\Theta \subset I\!R^d$. In unsupervised image segmentation, the goal is double: estimate the parameter of the model, and recover the hidden classification. The energy function H may depend on a parameter, which weights the local spatial dependencies. Its estimation is not an easy task and adds to the complexity of the problem. Throughout this paper, this parameter is supposed known and is not mentioned in the notation. Thus, the model parameter is θ.
The EM algorithm is a procedure widely used for parameter estimation when part of the data is missing. It aims at maximizing the log-likelihood of the model

$$L(\theta, \mathbf{y}) = \log P_G(\mathbf{y} \mid \theta), \tag{2}$$

by maximizing iteratively the conditional expectation of the complete log-likelihood. More precisely, if $\theta^{(q-1)}$ is the current estimate of the parameter at iteration $q-1$, this quantity is

$$Q(\theta \mid \theta^{(q-1)}) = I\!E_{P_G}[\log P_G(\mathbf{y}, \mathbf{Z} \mid \theta) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}]\,. \tag{3}$$

Assume that for all θ' in Θ, $Q(. \mid \theta')$ has a global maximum and that it is unique. Then, iteration q of the EM algorithm is divided in two steps:

(E)-step: Compute $Q(\theta \mid \theta^{(q-1)})$,

(M)-step: Update the current estimate of θ to

$$\theta^{(q)} = \arg\max_{\theta} Q(\theta \mid \theta^{(q-1)}).$$

Difficulties arise in the E step in the case of hidden Markov models. It involves the computation of the conditional probabilities $P_G(z_i \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)})$, which induces a sum over all the possible configurations of the field $\mathbf{Z}$. It is not tractable in general, and approximations are required. A solution is to use Monte Carlo Markov Chains (MCMC) methods to generate realizations of the target distribution, and then approximate the expectation (3) by the empirical mean. The corresponding procedure is called MCEM, (see Wei and Tanner 1990). Given the value $\theta^{(q-1)}_{T_{q-1}}$ of the parameter at iteration $q-1$, iteration q of MCEM is as follows:

(MC)-step: Generate T_q realizations $\mathbf{z}(q,1), \mathbf{z}(q,2), \ldots, \mathbf{z}(q,T_q)$ of an ergodic Markov chain whose invariant distribution is $P_G(\mathbf{z} \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)})$. Set

$$Q_{T_q}(\theta \mid \theta^{(q-1)}_{T_{q-1}}) = \frac{1}{T_q} \sum_{t=1}^{T_q} \log P_G(\mathbf{y}, \mathbf{z}(q,t) \mid \theta).$$

(M)-step: Update the current estimate of θ to

$$\theta_{T_q}^{(q)} = \arg\max_{\theta} Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}).$$

Note that if the spatial parameter had to be estimated as well, in the E step approximations of the probabilities $P_G(\mathbf{z} \mid \mathbf{Y} = \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}, \beta_{T_{q-1}}^{(q-1)})$ and $P_G(\mathbf{z} \mid \beta_{T_{q-1}}^{(q-1)})$ would be necessary, which increases the difficulty. We suppose $Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)})$ has a unique global maximum in θ, so that the M step is well defined. In practice, the number of iterations T_q can vary with q. Wei and Tanner (1990) recommend to start with small values, and increase T_q as the estimate $\theta_{T_q}^{(q)}$ gets closer to the maximizer. It is also possible to start by running the chain $\{\mathbf{z}(q,t)\}_{t\in \mathbb{N}}$ for a burn-in of m_0 iterations, until the stationary distribution is reached, and redefine Q_{T_q} as

$$Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) = \frac{1}{T_q} \sum_{t=m_0+1}^{m_0+T_q} \log P_G(\mathbf{y}, \mathbf{z}(q,t) \mid \theta).$$

A well known example of MCMC method which can be used in the MC step of MCEM is the Gibbs sampler (Geman and Geman 1984). It consists of running a Markov chain whose transition probabilities are given by the local characteristics of the distribution of interest $P_G(z_i \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}, \mathbf{Y} = \mathbf{y}, \theta^{(q-1)})$. Throughout this paper we will refer to this method for illustration. In our experiments the simulations of the MC step are carried out using the Gibbs sampler and details about its implementation are given in Section 5.

3 Related algorithms

In order to point out links between the MCEM algorithm and other stochastic versions of the EM procedure, we compare the estimate $\theta_{EM}^{(q)}$, obtained after one iteration of EM, with $\theta_{T_q}^{(q)}$, the one obtained after one iteration of MCEM, when considering the same current value $\theta_{T_{q-1}}^{(q-1)}$ to initialize both methods. Consider the vector of conditional probabilities,

$$\begin{aligned} \mathbf{p}_G^{(q)} &= \{P_G(Z_i = k \mid \mathbf{Y} = \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}),\ i \in S,\ k \in \{1,\ldots,K\}\} \\ &= \{p_{G\,i,k}^{(q)},\ i \in S,\ k \in \{1,\ldots,K\}\}. \end{aligned}$$

The function Q defined by (3) can be written

$$\begin{aligned} Q(\theta \mid \theta_{T_{q-1}}^{(q-1)}) &= \mathbb{E}_{P_G}[\log g(\mathbf{y} \mid \mathbf{Z},\ \theta) \mid \mathbf{Y} = \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}] \\ &\quad + \mathbb{E}_{P_G}[\log P_G(\mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}]. \end{aligned} \tag{4}$$

The second term on the right hand-side of equation (4) does not depend on θ. The first term, denoted $R(\theta \mid \mathbf{p}_G^{(q)})$ in the sequel, depends on $\theta^{(q-1)}$ only through $\mathbf{p}_G^{(q)}$, and can be further expanded, using (1),

$$R(\theta \mid \mathbf{p}_G^{(q)}) = \sum_{i\in S} \sum_{k=1}^{K} p_{G\,i,k}^{(q)} \log g_i(y_i \mid Z_i = k,\ \theta).$$

If $|S|$ is the number of elements in S and $\mathcal{P}$ is the set $\{\mathbf{p} \in [0,1]^{|S|\times K} / \forall i \in S, \sum_{k=1}^{K} p_{i,k} = 1\}$, we will denote by $\mathbf{\Phi}$ the function on $\mathcal{P}$, taking values in Θ, given by

$$\forall \mathbf{p} \in \mathcal{P},\ \mathbf{\Phi}(\mathbf{p}) = \arg\max_{\theta} R(\theta \mid \mathbf{p}),$$

and $\mathbf{\Phi}_{q:1}$ denotes q compositions of the function $\mathbf{\Phi}$. Then, one iteration of the EM algorithm, when the current value of the parameter is $\theta_{T_{q-1}}^{(q-1)}$, is equivalent to

$$\theta_{EM}^{(q)} = \mathbf{\Phi}(\mathbf{p}_G^{(q)}). \tag{5}$$

Let $\mathbf{d}$ be the indicator function in $\{0,1\}^{|S|\times K}$ with generic element

$$d_{i,k}(\mathbf{z}) = \begin{cases} 1, & \text{if } z_i = k \\ 0, & \text{if } z_i \neq k. \end{cases}$$

The vector of the empirical conditional frequencies is $\mathbf{f}^{(q)} = \{f_{i,k}^{(q)},\ i \in S,\ k \in \{1,\dots,K\}\}$, with

$$f_{i,k}^{(q)} = \frac{1}{T_q}\sum_{t=1}^{T_q} d_{i,k}(\mathbf{z}(q,t)). \tag{6}$$

In the MCEM algorithm, the function to be maximized in the E step is

$$Q_{T_q}(\theta \mid \theta_{T_{q-1}}^{(q-1)}) = \frac{1}{T_q}\sum_{t=1}^{T_q} \log g(\mathbf{y} \mid \mathbf{z}(q,t),\ \theta) + \frac{1}{T_q}\sum_{t=1}^{T_q} \log P_G(\mathbf{z}(q,t)).$$

Here again, the second term on the right hand-side of this equation does not depend on θ. The first term is equal to

$$\begin{aligned}\frac{1}{T_q}\sum_{t=1}^{T_q} \log g(\mathbf{y} \mid \mathbf{Z} = \mathbf{z}(q,t),\ \theta) &= \sum_{i\in S}\sum_{k=1}^{K} f_{i,k}^{(q)} \log g_i(y_i \mid Z_i = k,\ \theta) \\ &= R(\theta \mid \mathbf{f}^{(q)}).\end{aligned}$$

Finally,

$$\theta_{T_q}^{(q)} = \mathbf{\Phi}(\mathbf{f}^{(q)}). \tag{7}$$

Thus, the updating of θ by MCEM uses the same formula that in the E step of EM, with the conditional probabilities replaced by the empirical conditional frequencies. The MCEM algorithm for HMRFs, under the assumption that the spatial parameter is known and when the MC step is carried out using the Gibbs sampler, is thus equivalent to the EM/MPM algorithm of Comer and Delp (2000). The procedure is also closely related to the version B of the stochastic Restoration-Estimation algorithm (SRE$_B$) of Qian and Titterington (1991) whose iteration is defined by a Restoration step, similar to the MC step of MCEM, and an Estimation step defined by

$$\theta_{SRE_B}^{(q)} = \arg\max_{\theta} \log P(\mathbf{y} \mid \frac{1}{T_q}\sum_{t=1}^{T_q} \mathbf{z}(q,t), \theta).$$

Since, for MCEM

$$\theta_{T_q}^{(q)} = \arg\max_{\theta} \frac{1}{T_q}\sum_{t=1}^{T_q} \log P(\mathbf{y} \mid \mathbf{z}(q,t), \theta),$$

MCEM and SRE$_B$ represent the same procedure when the function $\log P(\mathbf{y} \mid \mathbf{z}, \theta)$ is linear in the variable $\mathbf{z}$. The underlying idea in the MCEM algorithm for HMRFs is to use MCMC methods to get an approximation $\hat{\mathbf{p}}^{(q)}$ of the intractable conditional probabilities $\mathbf{p}_G^{(q)}$ and replace them in the EM formula. It consists of the empirical frequencies for this algorithm, but other choices can be considered and the method generalized. These stochastic versions of EM (including MCEM) will be referred to as MCEM-like algorithms. For instance, the vector $\mathbf{p}_G^{(q)}$ can be approximated in the MC step using $\hat{\mathbf{p}}^{(q)}$ given by

$$\forall\ i \in S, \forall\ k \in \{1,\dots,K\},\ \hat{p}_{i,k}^{(q)} = \frac{1}{T_q}\sum_{t=1}^{T_q} P_G(Z_i = k \mid \mathbf{Z}_{N(i)} = \mathbf{z}_{N(i)}(q,t), \mathbf{Y} = \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}). \tag{8}$$

The nature of this approximation is different than the MCEM approximation (6). If $T_q = 1$, $\mathbf{f}^{(q)}$ reduces to a Dirac distribution, giving probability one to realization $\mathbf{z}(q,1)$, which is not the case for $\hat{\mathbf{p}}^{(q)}$. More specifically, the approximation of $p^{(q)}_{G\,i,k}$ defined by (8) is obtained by approximating the influence of the neighborhood of pixel i by the simulated value $\mathbf{z}_{N(i)}(q,1)$. The case $T_q = 1$ has been studied in Celeux, Forbes, and Peyrard (2001), using the Gibbs sampler in the MC step. The corresponding algorithm is called the Simulated Field algorithm. When $T_q \geq 1$, this procedure will be referred to as the Generalized Simulated Field (GSF) algorithm. Convergence results for the MCEM and the GSF algorithms are given in the next section, and their behavior is compared in Section 5.

4 Theoretical results about MCEM-like algorithms

In this section, two convergence aspects of the MCEM-like algorithms are studied. Firstly, we consider the random perturbation $\theta^{(q)}_{T_q} - \theta^{(q)}_{EM}$ added at iteration q of MCEM to the EM estimate, due to the stochastic approximation of the function Q. If this difference is small, an MCEM iteration behaves as an EM iteration. In particular, the non decreasing of the log-likelihood is a well known property of the Generalized EM (GEM) procedures (Dempster, Laird, and Rubin 1977) which is the key of their appealing behavior. The theory available about GEM algorithms is used to establish results on the convergence of the MCEM parameter estimate, and more generaly to the estimate of a MCEM-like algorithm.

4.1 Convergence of the log-likelihood

Theorem 4.1 *If the following properties are satisfied*

(i) $R(.\mid \mathbf{p})$ *is unimodal, whatever* $\mathbf{p}$ *in* $\mathcal{P}$*;*

(ii) $Q(.\mid \theta')$ *is a continuous function on* Θ*, for all* $\theta' \in \Theta$*;*

(iii) $\mathbf{\Phi}(.)$ *is a continuous function on* $\mathcal{P}$*;*

(iv) *if* $P_G(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y}, \theta) = P_G(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y}, \theta')$ *almost everywhere, then* $\theta = \theta'$*;*

then,

$$P(\forall q \in I\!N,\ \exists T^* \in I\!N /\ \forall T_q > T^*, \quad L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) \geq 0) = 1.$$

If in addition,

(v) *the log-likelihood* $L(\theta, \mathbf{y})$ *is bounded above on* Θ*;*

with probability one it exists a sequence $\{T_q\}_{q\in I\!N}$ *such that the sequence of log-likelihoods* $\{L(\theta^{(q)}_{T_q}, \mathbf{y})\}_{q\in I\!N}$ *generated by the MCEM algorithm converges.*

Proof. Property (i) implies that the function $\mathbf{\Phi}(.)$ is well defined, as well as the M step of EM and the MC step of MCEM. The ergodicity of the chain $\{\mathbf{z}(q,t)\}_{1\leq t\leq T_q}$ generated at iteration q implies that for every function h on $\{1,\ldots,K\}^{|S|}$,

$$P\left(\lim_{T_q\to\infty} \frac{1}{T_q}\sum_{t=1}^{T_q} h(\mathbf{z}(q,t)\,) = I\!E_{P_G}[h(\mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}}]\right) = 1. \tag{9}$$

See for instance Geman and Geman (1984) (Theorem C) for a proof in the case of a chain generated by the Gibbs sampler. Applying (9) for $h = \mathbf{d}$ leads to,

$$P(\lim_{T_q\to\infty} \mathbf{f}^{(q)} = I\!E_{P_G}[\mathbf{d}(\mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}}]) = 1.$$

As

$$I\!E_{P_G}[d_{i,k}(\mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}}] = p^{(q)}_{G,ik},$$

we have

$$P(\lim_{T_q \to \infty} \mathbf{f}^{(q)} = \mathbf{p}_G^{(q)}) = 1. \tag{10}$$

and since $\boldsymbol{\Phi}$ is a continuous function,

$$P(\lim_{T_q \to \infty} \boldsymbol{\Phi}(\mathbf{f}^{(q)}) = \boldsymbol{\Phi}(\mathbf{p}_G^{(q)})) = 1.$$

According to (5) and (7), it means that with probability one the MCEM estimate converges towards the estimate resulting of one iteration of EM with initial value $\theta^{(q-1)} = \theta_{T_{q-1}}^{(q-1)}$,

$$P(\lim_{T_q \to \infty} \theta_{T_q}^{(q)} = \theta_{EM}^{(q)}) = 1. \tag{11}$$

We consider now the evolution of $L(\theta_{T_q}^{(q}, \mathbf{y})$ after one iteration of MCEM Using now the continuity of $Q(. \mid \theta_{T_{q-1}}^{(q-1)})$, (11) implies that

$$P(\lim_{T_q \to \infty} Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) = Q(\theta_{EM}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)})) = 1. \tag{12}$$

Consider the event $A_q = \{\forall \epsilon > 0 \ \exists T^* \ / \ \forall T_q > T^*, \ |Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{EM}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)})| < \epsilon\}$. By definition of $\theta_{EM}^{(q)}$, the difference $\alpha = Q(\theta_{EM}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{T_{q-1}}^{(q-1)} \mid \theta_{T_{q-1}}^{(q-1)})$ is non negative. The two cases $\alpha > 0$ and $\alpha = 0$ must be treated separately. Firstly, if $\alpha > 0$, it exists $\epsilon_0 > 0$ such that $\alpha - \epsilon_0 \geq 0$. If the event A_q is observed then,

$$\exists T^0 \ / \ \forall T_q > T^0, \ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{EM}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) > -\epsilon_0,$$

and for this value T^0, we have

$$\forall T_q > T^0, \ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{EM}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) + \alpha > \alpha - \epsilon_0.$$

Then,

$$\forall T_q > T^0, \ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{T_{q-1}}^{(q-1)} \mid \theta_{T_{q-1}}^{(q-1)}) \geq 0.$$

The likelihood (2) of the model can be rewritten,

$$L(\theta, \mathbf{y}) = Q(\theta \mid \theta_{T_{q-1}}^{(q-1)}) - H(\theta \mid \theta_{T_{q-1}}^{(q-1)}),$$

where

$$H(\theta \mid \theta_{T_{q-1}}^{(q-1)}) = \mathbb{E}_{P_G}[\log P_G(\mathbf{Z} \mid \mathbf{Y} = \mathbf{y}, \theta) | \mathbf{Y} = \mathbf{y}, \theta_{T_{q-1}}^{(q-1)}].$$

It can be proved, using Jensen's inequality, that H is a function of θ which has a global maximum at $\theta = \theta_{T_{q-1}}^{(q-1)}$, (see Dempster, Laird, and Rubin 1977). Then it comes

$$\forall T_q > T^0, \ L(\theta_{T_q}^{(q)}, \mathbf{y}) - L(\theta_{T_{q-1}}^{(q-1)}, \mathbf{y}) \geq 0. \tag{13}$$

We turn now to the case $\alpha = 0$. Assumption (i) implies that $Q(. \mid \theta_{T_{q-1}}^{(q-1)})$ is unimodal. Thus, if $\alpha = 0$, the two estimates $\theta_{T_{q-1}}^{(q-1)}$ and $\theta_{EM}^{(q)}$ are equal. Consider the difference $\delta \geq 0$ defined by

$$\delta = H(\theta_{T_{q-1}}^{(q-1)} \mid \theta_{T_{q-1}}^{(q-1)}) - H(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}).$$

If $\delta > 0$, then, for a given $\epsilon_1 > 0$ such that $\delta - \epsilon_1 \geq 0$, A_q implies that

$$\exists T^1 \ / \ \forall T_q > T^1, \ Q(\theta_{T_q}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) - Q(\theta_{EM}^{(q)} \mid \theta_{T_{q-1}}^{(q-1)}) > -\epsilon_1,$$

and since $\theta^{(q-1)}_{T_{q-1}} = \theta^{(q)}_{EM}$,

$$\forall T_q > T^1,\ Q(\theta^{(q)}_{T_q} \mid \theta^{(q-1)}_{T_{q-1}}) - Q(\theta^{(q-1)}_{T_{q-1}} \mid \theta^{(q-1)}_{T_{q-1}}) > -\epsilon_1.$$

Then,

$$\forall T_q > T^1,\ L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) > -\epsilon_1 + \delta \geq 0. \tag{14}$$

Now, Dempster, Laird, and Rubin (1977) demonstrated that the difference δ is equal to zero if and only if $P_G(\mathbf{z} \mid \mathbf{Y} = \mathbf{y}, \theta^{(q)}_{T_q}) = P_G(\mathbf{z} \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}})$ almost everywhere. We deduce from assumption (iv), that in that case, $\theta^{(q)}_{T_q} = \theta^{(q-1)}_{T_{q-1}}$. Then, $L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) = 0$.
Finally, combining (13) and (14), whatever $\alpha \geq 0$,

$$\exists T^* = \sup(T^0, T^1)\ /\ \forall T_q > T^*, \quad L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) \geq 0, \tag{15}$$

with T^* finite, as T^0 and T^1. The event B_q defined by (15) is implied by A_q, so that, by (12)

$$\forall q \in I\!N,\ P(B_q) \geq P(A_q) \quad = \quad 1,$$

and since the intersection is countable,

$$P(\bigcap_q B_q) \quad = \quad 1,$$

In terms of the log-likelihood's evolution, it means that

$$P(\forall q \in I\!N,\ \exists T^* \in I\!N /\ \forall T_q > T^*, \quad L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) \geq 0) = 1.$$

If in addition, the log-likelihood $L(\theta, \mathbf{y})$ is bounded above, this result implies that almost surely it exists a sequence $\{T_q\}_{q \in I\!N}$, such that the sequence $\{L(\theta^{(q)}_{T_q}, \mathbf{y})\}_{q \in I\!N}$ converges.

□

Some remarks are in order about Theorem 4.1. Firstly, the result can be generalized to any MCEM-like algorithm, as soon as the corresponding approximation $\hat{\mathbf{p}}^{(q)}$ satisfies

$$P(\lim_{T_q \to \infty} \hat{\mathbf{p}}^{(q)} = \mathbf{p}^{(q)}_G) = 1. \tag{16}$$

For such a procedure, the proof of Theorem 4.1 is still valid, (10) being replaced by (16). In particular, (16) is satisfied by the GSF algorithm: the distribution $\hat{\mathbf{p}}^{(q)}$ defined by (8) converges towards

$$I\!E_{P_G}[P_G(Z_i = k \mid \mathbf{Z}_{N(i)}, \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}] \quad = \quad P_G(Z_i = k \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}),$$

according to the ergodic property (9).

In Theorem 4.1, assumption (v) can be replaced by the stronger assumption $(v)'$ *the g_i's are positive and bounded uniformly in θ*. Then assumptions (i), (ii), (iv) and $(v)'$ are sufficient to ensure that almost surely the sequence of log-likelihoods is not decreased and converges. Condition (iii) is no more required. Indeed, we have the following lemma

Lemma 4.1 *If the g_i's are positive and bounded uniformly in Θ,*

$$P(\lim_{T_q \to \infty} Q_{T_q}(\theta \mid \theta^{(q-1)}_{T_{q-1}}) = Q(\theta \mid \theta^{(q-1)}_{T_{q-1}}),\ \text{uniformly in } \theta) = 1.$$

It ensures that the property (11) is still true, since the uniform convergence of a sequence of functions $\{a_n\}_{n\in\mathbb{N}}$ towards a function a implies the convergence of $\{\arg\max_x a_n(x)\}_{n\in\mathbb{N}}$ towards $\arg\max_x a(x)$. The rest of the proof of Theorem 4.1 remains the same.

Proof of Lemma 4.1. Assumption (v') implies that it exists a positive constant, M, such that,

$$\forall i \in S,\ \forall k \in \{1,\ldots,K\}, \quad |\log g_i(y_i \mid Z_i = k,\ \theta)\ | \le M.$$

Consider the difference $D^{(q)}(\theta) = |\ Q_{T_q}(\theta \mid \theta^{(q-1)}_{T_{q-1}}) - Q(\theta \mid \theta^{(q-1)}_{T_{q-1}})\ |$.

$$\begin{aligned} D^{(q)}(\theta) \quad \le \quad & |\ \frac{1}{T_q}\sum_{t=1}^{T_q} \log P_G(\mathbf{y} \mid \mathbf{Z} = \mathbf{z}(q,t), \theta) - \mathbb{E}_{P_G}[\log P_G(\mathbf{y} \mid \mathbf{Z}, \theta) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}}]\ | \\ & + |\ \frac{1}{T_q}\sum_{t=1}^{T_q} \log P_G(\mathbf{z}(q,t)) - \mathbb{E}_{P_G}[\log P_G(\mathbf{Z}) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}}]\ |. \end{aligned} \tag{17}$$

The first term, $D_1^{(q)}(\theta)$, on the right hand-side of inequality (17), can be bounded above uniformly in θ. If $\mathbb{E}_{\mathbf{f}^{(q)}}$ denotes the expectation under the distribution defined by $\mathbf{f}^{(q)}$,

$$\begin{aligned} \forall \theta \in \Theta, \quad D_1^{(q)}(\theta) \quad & = \quad |\ \mathbb{E}_{\mathbf{f}^{(q)}}[\log P_G(\mathbf{y} \mid \mathbf{Z}, \theta)] - \mathbb{E}_{P_G}[\log P_G(\mathbf{y} \mid \mathbf{Z}, \theta) \mid \mathbf{Y} = \mathbf{y}, \theta^{(q-1)}_{T_{q-1}}] \\ & \le \quad M \sum_{i\in S}\sum_{k=1}^{K} |\ f^{(q)}_{i,k}\ -\ p^{(q)}_{G\,i,k}\ |. \end{aligned}$$

Since by (10) the empirical frequencies, $\mathbf{f}^{(q)}$, converge almost surely towards the theoretical ones, $\mathbf{p}_G^{(q)}$,

$$P(\lim_{T_q\to\infty} \sup_\theta D_1^{(q)}(\theta) = 0) \quad \ge \quad P(\forall i \in S,\ \forall k \in \{1,\ldots,K\},\ \lim_{T_q\to\infty} |\ p^{(q)}_{G\,i,k} - f^{(q)}_{i,k}\ | = 0) = 1.$$

The second term on the right hand-side of (17), denoted $D_2^{(q)}$, is independent on θ. Applying property (9) with $h(\mathbf{z}) = \log P_G(\mathbf{z})$ leads to

$$P(\lim_{T_q\to\infty} D_2^{(q)} = 0) = 1.$$

Finally, $P(\lim_{T_q\to\infty} \sup_\theta D^{(q)}(\theta) = 0) = 1$, which proves the almost sure uniform convergence.

□

In image analysis, the observed data $\mathbf{y}_i$ is often supposed arising from a univariate Gaussian density conditionally on $\mathbf{z}_i$. In that case, if one of the observation y_i happens to be equal to the mean m_k of one component, and the variance σ_k^2 tends to zero, then the density $g_i(y_i \mid Z_i = k,\ \theta)$ is not finite. By constraining the parameter space, for instance assuming for all $k \in \{1,\ldots,K\}$, $\sigma_k \in [\underline{\sigma}, \overline{\sigma}]$ with $\underline{\sigma} > 0$, this problem is suppressed and Lemma 4.1 can be used.

4.2 Convergence of the estimate

The previous section establishes conditions for the almost sure convergence of the sequence of log-likelihoods towards some L^*, in MCEM-like procedures. We consider now the issues of the nature of L^* and of the convergence of the sequence $\{\theta^{(q)}_{T_q}\}_{q\in\mathbb{N}}$ to a stationary point or a local maximum of $L(\theta, \mathbf{y})$. Our results are derived from the study of the convergence properties of the GEM algorithms by Wu (1983). Dempster, Laird, and Rubin (1977) defined a GEM algorithm as a less demanding algorithm than EM, for which the M step requires only

$$Q(\theta^{(q)} \mid \theta^{(q-1)}) \quad \geq \quad Q(\theta^{(q-1)} \mid \theta^{(q-1)}).$$

For any sequence $\theta^{(q)}$ of a GEM algorithm, we have

$$L(\theta^{(q)}, \mathbf{y}) \geq L(\theta^{(q-1)}, \mathbf{y}),$$

which ensures the algorithm inherits most of the EM convergence properties. In the sequel we adopt Wu's notation. Let $\mathcal{S}$ be the set of stationary points, and $\mathcal{M}$ the set of local maxima of $L(\theta, \mathbf{y})$. A direct application of Theorem 1 of Wu (1983) leads to the following theorem,

Theorem 4.2 *Suppose that almost surely,*

(i) $\forall q \in I\!N$, $\exists T^* \,/\, \forall T_q > T^*$, $L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) \geq 0$;

(ii) $\forall q \in I\!N$, $\forall \theta^{(q-1)}_{T_{q-1}} \notin \mathcal{S}$ (*resp.*$\mathcal{M}$), $\exists T^{**} \,/\, \forall T_q > T^{**}$, $L(\theta^{(q)}_{T_q}, \mathbf{y}) - L(\theta^{(q-1)}_{T_{q-1}}, \mathbf{y}) > 0$;

(iii) *the function* $\arg\max\limits_{\theta} Q_{T_q}(\theta \mid \theta')$ *is closed over the complement of* $\mathcal{S}$ (*resp.* $\mathcal{M}$);

(iv) *the sequence* $\{L(\theta^{(q)}_{T_q}, \mathbf{y})\}_{q \in I\!N}$ *is bounded above for any* $\theta^{(0)}$ *in* Θ.

If for each iteration of a MCEM-like algorithm, T_q *is chosen such that* $T_q > \max(T^*, T^{**})$*, then almost surely, all the limits points of* $\{\theta^{(q)}_{T_q}\}_{q \in I\!N}$ *are stationary points (resp. local maxima) of* $L(\theta, \mathbf{y})$*, and* $L(\theta^{(q)}_{T_q}, \mathbf{y})$ *converges monotically to* $L^* = L(\theta^*, \mathbf{y})$ *for some* θ^* *of* $\mathcal{S}$ *(resp.* $\mathcal{M}$*).*

We have shown, with Theorem 4.1, conditions ensuring that assumption (i) is satisfied. This assumption means that with probability one the sequence of log-likelihoods in a MCEM-like algorithm behaves like in a GEM algorithm. This is of high interest since the non decreasing of the log-likelihood is the main assumption to establish results about the convergence of the GEM estimate. As a consequence, as soon as assumption (i) of Theorem 4.2 is satisfied, the estimate provided by a MCEM-like algorithm has the same properties than the GEM estimate, convergence been replaced by almost sure convergence. Theorem 4.2 gives an example of this link between GEM and MCEM asymptotic behavior. Conditions for almost sure pointwise convergence can also be derived from the theoretical results of Wu (1983) about the GEM procedure.

These results add to the ones established by Comer and Delp (2000) about the MCEM (or equivalently EM/MPM) convergence in the case where the spatial parameter is known. In their study, they do not use the fact that MCEM is related to GEM procedures and consequently do not take advantage of the available convergence results for the GEM algorithms. Their approach consists of comparing the estimate resulting of q iterations of MCEM, $\theta^{(q)}_{T_q} = \mathbf{\Phi}_{q:1}(\mathbf{f}^{(0)})$, with the estimate obtained after q iterations of EM, $\mathbf{\Phi}_{q:1}(\mathbf{p}^{(0)}_G)$. Their results guarantee the existence of an integer N and a sequence $\{T_q\}_{1 \leq q \leq N}$ such that with probability 1, $\theta^{(q)}_{T_q}$ can be made arbitrary close to the EM estimate of θ, when the MCEM and EM procedures are initialized with the same value. It is actually an asymptotic result in the sense that the number of iterations needed, T_q, can be infinite for all q, $1 \leq q \leq N$. The property established suggests at best that the MCEM estimate can converge in probability towards the maximizer of the log-likelihood of the model, while Theorem 4.2 exhibits conditions for almost sure convergence. The main difference in our approach and the one of Comer and Delp (2000) lies in the level of comparison with the EM algorithm. The proof of Theorem 4.1 is based on the control of the local error $\theta^{(q)}_{T_q} - \theta^{(q)}_{EM}$, instead of studying the global error $\mathbf{\Phi}_{q:1}(\mathbf{f}^{(0)}) - \mathbf{\Phi}_{q:1}(\mathbf{p}^{(0)}_G)$ as in Comer and Delp (2000).

5 Illustration

In this section the behavior of the MCEM and the GSF algorithms is illustrated on a hidden Markov model often used in image segmentation. The model for the hidden data $\mathbf{Z}$ is the K-color Potts model, defined by

$$P_G(\mathbf{z} \mid \beta) \quad = \quad W(\beta)^{-1} \exp(\beta \sum_{i \sim j} I_{\{z_i = z_j)\}}) \,, \tag{18}$$

where the notation $i \sim j$ represents any couple (i,j) of neighboring sites, and I is the indicator function, equal to 1 if $z_i = z_j$, and to zero otherwise. A second-order neighborhood system (*i.e.* the eight closer neighbors for each pixel) is considered. The Gibbs sampler is used to perform the MC step of the MCEM and the GSF algorithms. More specifically, only one site of S is visited at each step, in increasing order of the indices, so that $\mathbf{z}(q,t)$ and $\mathbf{z}(q,t+1)$ can only differ by one component, let say z_{i_t}. The transition probabilities are given by the full conditional probabilities of the Markov model,

$$\begin{aligned} & P(\mathbf{Z}(q,t+1) = \mathbf{z} \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z}(q,t) = \mathbf{z}', \theta^{(q-1)}_{T_{q-1}}) \\ & = \begin{cases} 0, \text{ } if \ \exists i \neq i_t \,/\, z_i \neq z'_i, \\ P_G(Z_{i_t} = z_{i_t} \mid \mathbf{Y} = \mathbf{y}, \mathbf{Z}_{N(i_t)} = \mathbf{z}'_{N(i_t)}, \theta^{(q-1)}_{T_{q-1}}) \text{ otherwise.} \end{cases} \end{aligned}$$

We suppose the conditional density of Y_i, given z_i is in class k, is gaussian,

$$g_i(y_i \mid Z_i = k, \theta) \; = \; \frac{1}{\sigma_k\sqrt{2\pi}} \exp\{-\frac{(y_i - m_k)^2}{2\sigma_k^2}\} \; ,$$

with $\theta = \{(m_k, \sigma_k), k = 1, \ldots, K\}$, m_k and σ_k^2 denoting the mean and variance of the density in the class k. The function $R(\theta \mid \mathbf{p}_G^{(q)})$ to be maximized in the E step of EM is equal to

$$R(\theta \mid \mathbf{p}_G^{(q)}) \;\; = \;\; \sum_{i\in S}\sum_{k=1}^{K} p^{(q)}_{G\,i,k} \log\left[\frac{1}{\sigma_k\sqrt{2\pi}}\exp\{-\frac{(y_i - m_k)^2}{2\sigma_k^2}\}\right] .$$

We experimented the procedures on the original 4-color image (a) of Figure 1. After addition of Gaussian noise, we obtained the image (b) of Figure 1. In this example the means m_k of the conditional distributions of the observed data are equal to k for $k \in \{1, \ldots, K\}$ and the standard deviations are the same for each class, $\sigma_k = 0.5$ for $k \in \{1, \ldots, K\}$. The parameter β of the Potts distribution is supposed equal to 1, and not estimated. Each algorithm is initialized with the same classification (Fig. 1 (c)), derived by partitioning the range of the pixel values in the noisy image in four regular intervals. Both the MCEM and the GSF algorithms are run with three different values of T_q: 1, 10, and 100 steps of the Gibbs sampler at each iteration. The number of iterations is set to $N = 100$, since no significant improvement was observed with more iterations for this example. The estimates of the model parameters provided by the different algorithms are given in Table 1, and the corresponding segmentations of the noisy image are shown in Figures 1 (d) to (i). A restoration $\hat{\mathbf{z}}$ is obtained by maximization of the approximation of the conditional probabilities,

$$\begin{aligned} \text{MCEM :} \qquad & \forall i \in S, \quad \hat{z}_i = \arg\max_k \hat{f}^{(N)}_{i,k}. \\ \text{GSF :} \qquad & \forall i \in S, \quad \hat{z}_i = \arg\max_k \hat{p}^{(N)}_{i,k}. \end{aligned}$$

The quality of the restorations are compared through the percentage of misclassified pixels (error rate), shown in Table 1. For this example, it appears that it is not necessary to choose large values for T_q. If too few iterations of the Gibbs sampler are carried out, the algorithms have not yet converged. But a choice of $T_q = 10$ already leads to good performances for both the MCEM and the GSF algorithms.

6 Conclusion

This work contributes to the analysis of the MCEM algorithm for HMRFs by showing that under suitable regularity conditions this procedure almost surely inherits the convergence properties of a GEM algorithm. We propose sufficient conditions to ensure that, with probability 1, the sequence of MCEM log-likelihoods is not decreased, so that the well known property of a GEM algorithm is recovered almost surely. We

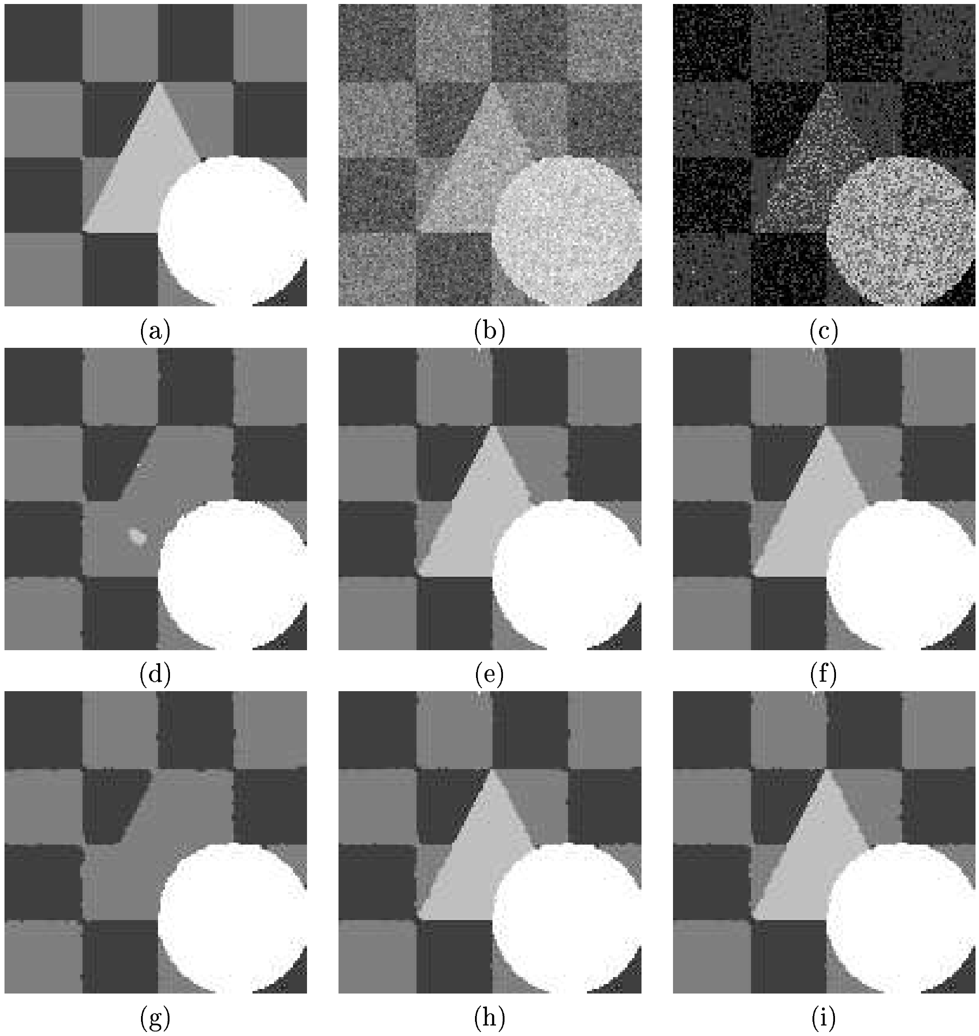

Figure 1: A 4-color image degraded with noise. (a) original image, (b) noisy image, (c) initial segmentation, (d), (e), (f) MCEM segmentations for $T_q = 1, 10, 100$, (g), (h), (i) GSF segmentations for $T_q = 1, 10, 100$.

	T_q	μ_1	μ_2	μ_3	μ_4	σ_1	σ_2	σ_3	σ_4	error rate
true value	-	1	2	3	4	0.5	0.5	0.5	0.5	-
MCEM	1	1.0	2.2	3.5	4.0	0.5	0.6	0.6	0.5	9.9
MCEM	10	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.6	0.5
MCEM	100	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.5
GSF	1	1.0	2.0	2.2	4.0	0.5	0.3	0.6	0.5	10.3
GSF	10	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.5
GSF	100	1.0	2.0	3.0	4.0	0.5	0.5	0.5	0.5	0.4

Table 1: Parameter estimates and error rates for the MCEM and the GSF algorithms, for the 4-color image of Figure 1 (a).

derive from the GEM theory sufficient conditions under which the MCEM estimate almost surely converges towards a maximizer of the log-density of the observations, which validate the procedure. The advantage of these results is that they can easily be extended to more general stochastic versions of EM based on a plug-in method. In this family, we studied in particular the Generalization of the so-called Simulated Field algorithm (GSF). Numerical experiments on the MCEM and the GSF algorithms, for a simple example of image segmentation, suggest that both procedures perform equivalently and that the asymptotic convergence stated by the theoretical results can be reached with a reasonable number of iterations for both the Gibbs sampler and the algorithm.

References

Celeux, G., F. Forbes, and N. Peyrard (2001). EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation. Research Report 4105, Inria, available at http://www.inria.fr/rrrt/rr-4105.html.

Comer, M. L. and E. J. Delp (2000). The EM/MPM algorithm for segmentation of textured images: analysis and further experimental results. *IEEE Transactions on Image Processing 9*(10), 1731–1744.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Geman, S. and D. Geman (1984). Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Qian, W. and D. M. Titterington (1991). Estimation of parameters in hidden Markov models. *Philosophical Transactions of the Royal Society of London, Series A 337*, 407–428.

Wei, G. and M. A. Tanner (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association 85*(411), 699–704.

Wu, C. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics 11*, 95–103.

Zhang, J. (1992). The Mean Field Theory in EM Procedures for Markov Random Fields. *IEEE, Transactions on Signal Processing 40*(10), 2570–2583.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399